Collaborative Multicast Beamforming for Content Delivery by Cache-enabled Ultra Dense Networks

Nguyen, H. T., Tuan, H. D., Duong, Q., Poor, H. V., & Hwang, W-J. (2019). Collaborative Multicast Beamforming for Content Delivery by Cache-enabled Ultra Dense Networks. *IEEE Transactions on Communications*, *67*(5), 3396 - 3406. https://doi.org/10.1109/TCOMM.2019.2894797

Published in:
IEEE Transactions on Communications

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2019 IEEE. This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date:06. May. 2023

Collaborative Multicast Beamforming for Content Delivery by Cache-enabled Ultra Dense Networks

H. T. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and W-J. Hwang

***Abstract*—Caching and multicast have surged as an effective tool to alleviate the heavy load from the backhaul links while enabling the content-centric delivery in communication networks. The main focus was about cache placements to manage the network delay and backhaul transmission cost. An important issue of optimizing the cost efficiency in content delivery has not been addressed. The paper tackles this issue by proposing collaborative multicast beamforming at cache-enabled ultra-dense networks. Our objective is to maximize the cost efficiency, which is defined as the ratio of the content throughput to the sum of power consumption and backhaul cost, in providing the quality-of-service for content delivery. Zero-forcing beamforming and generalized zero-forcing beamforming are employed to force the multi-content interference to zero or mitigate it while amplifying the desired signals for users. These problems of the collaborative multicast beamforming design are computationally difficult. We develop path-following algorithms, which invoke a simple convex quadratic program at each iteration, for their solution. Numerical results are provided to demonstrate the computational efficiency of the proposed algorithms and also give insights into the impact of caching on the cost efficiency.**

***Index Terms*—Content-centric communications, caching, multicast, collaborative beamforming, cost efficiency, quality-of-service constraint, non-convex optimization, path-following algorithm.**

I. Introduction

In face of the quantum increased communication demand from social networking platforms, the global mobile data traffic is expected to exponentially grow in the 5th generation cellular networks (5G) [1], [2]. Content-centric communications [3] have been introduced to support the content-service, which is one of the 5G's features. Multicast for content items that are concurrently requested by users is a practical way to save the data traffic in content-service communications [4]. Multicast beamforming has been addressed in [5], [6].

This work was supported in part by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2018-2016-0-00318) supervised by the IITP(Institute for Information & Communications Technology Promotion), in part by Institute for Computational Science and Technology, Hochiminh city, Vietnam, and in part by the U.K. Royal Academy of Engineering Research Fellowship under Grant RF1415\14\22

Huy T. Nguyen and Won-Joo Hwang are with the Department of Information and Communication System, Inje University, 621-749, Gimhae, Gyeongnam, Korea. (email: huynguyencse@gmail.com, ichwang@inje.ac.kr).

Hoang Duong Tuan is with the School of Electrical and Data Engineering, University of Technology Sydney, Broadway, NSW 2007, Australia (e-mail: tuan.hoang@uts.edu.au).

Trung Q. Duong is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K. (e-mail: trung.q.duong@qub.ac.uk).

H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA (e-mail: poor@princeton.edu)

Deployment of spatially distributed small base stations (SBSs) is a 5G tool for increasing the network throughput or enhancing the throughput at the cell-edge users. However, the promises cannot hold during the high-peak traffic time due to the bandwidth limitation of the backhaul links, which are the backbone links between SBSs and the core network [7], [8]. By equipping SBSs with local storage of high processing rate [9], caching at the SBSs is considered as an effective way to alleviate the burden on the backhaul links [8], [10].

Mixed caching and multicast have been discussed in [11]–[13] in the context of heterogeneous networks of SBSs and macro base stations (MBS) serving multiple users. A scheme of unicast transmissions for the disperse popular content items that are cached at SBSs and a multicast transmission by an MBS for the concurrently requested content items was proposed in [11]. A network of an MBS, which stores all content items, and SBSs, which store a limited number of content items was considered in [12]. A stochastic content multicast scheduling was proposed to minimize the average network delay and power costs when the MBS and the SBSs operate nonconcurrently. By taking into account the backhaul constraints, [13] examined the successful transmission probability in a large-scale heterogeneous network of MBSs storing the same content items and SBSs storing different content items randomly.

As the existing works on caching were mainly interested in alleviating the burden on the backhaul links or shortening the transmission delay, only a few works have studied the impact of caching on energy efficiency (EE), which is one of the important performance metrics in 5G [14]. For instance, the relationship between the EE and cache hit ratio in user-centric networks was investigated in [15]. Some key factors that contribute to the EE gain by caching have been identified in [16]. An EE-aware multicast beamforming in the context of cache-enabled cloud radio access networks was considered [17], which requires a high computational cost with no guaranteed optimality.

The present paper considers content-centric communications by ultra-dense networks (UDNs) [18], [19], which consist of the large numbers of cache-enabled spatially distributed SBSs. The delivery of the cached content items is expected to be cost-efficient since there is no need to fetch them from the core network. In contrast, the central process needs to fetch the cache-missed content items to SBSs through the backhaul links, which are not only costly but also limited. The users form disjoint clusters based on their requests, i.e. those that request the same content belong to the same cluster. On the other hand, those SBSs that have the same content

at their disposal also form a collaborative multicast group to deliver this content via multicast transmission. Thus, each SBS can belong to several multicast groups and each users' cluster can be served by multiple SBSs. As a result, our challenges are dealing with diverse interferences such as SBS interference and multi-content interference. We consider the problem of collaborative multicast beamforming to maximize the network cost efficiency, in terms of the ratio of the sum content throughput and the sum of the consumption power and backhaul cost for fetching the cache-missed contents, under a quality-of-service (QoS) constraint in terms of their throughput thresholds. To achieve computational tractability, beamformers are sought in the class of zero-forcing (ZF) beamforming or generalized zero-forcing (GZF) beamforming[1]. Unlike the user throughput in user-centric communication, which is obviously a smooth function of beamforming vectors (see e.g. [24]–[26] and [27]), the content throughput is no longer a smooth function. Consequently, such beamforming designs are modeled as nonconvex and nonsmooth optimization problems, which are computationally difficult. We aim to propose path-following computational procedures of low-complexity computation, which invoke a simple convex quadratic program in each iteration, for their solution.

The rest of this paper is organized as follows. The problem statement is presented in Section II. Collaborative multicast ZF beamforming is developed in Section III while collaborative multicast GZF beamforming is developed in Section IV. Numerical results are presented in Section V and conclusions are drawn in Section VI.

Notation: Boldface upper denote matrices, bold lower-case letters denote column vectors, and lower-case letters denote scalars, respectively. The Hermitian transpose operator is represented by $(.)^H$. $\|.\|$ and $|.|$ denote the Euclidean norm and the absolute value of a complex scalar, respectively. The notation $\Re\{.\}$ denotes the real part of a complex number. A Gaussian vector $\mathbf{x}$ with mean μ and covariance σ_n^2 is denoted by $\mathbf{x} \sim \mathcal{CN}(\mu, \sigma_n^2)$. $|A|$ for the set A is its cardinality.

II. Problem Statement

Consider an ultra-dense downlink network consisting of N_{BS} spatially distributed N_t-antenna SBSs indexed by $\mathcal{Q} = \{1, \ldots, N_{BS}\}$ to serve N_u single-antenna users (UEs) as illustrated in Fig. 1. Each SBS's cache stores N_s contents. There are a total of N_c contents, which are either distributed among SBS caches or are stored in the database at the central processor (CP). The capacity size of all content items is $\mathcal{C}_f$. Each SBS cannot store all N_c contents, so $N_s < N_c$ [28]. Denote by $\mathcal{F}_q$, $q \in \mathcal{Q}$ the set of the content items in SBS q's cache. Thus, it is true that $|\mathcal{F}_q| \leq N_s, \forall q \in \mathcal{Q}$.

Each UE requests only one content at each interval time. The content popularity is distributed according to a Zipf-like distribution [29], in which the requesting probability of the f-th content item is defined by

$$p_f = \frac{f^{-\delta}}{\sum_{j=1}^{N_f} j^{-\delta}}.$$

[1]ZF beamforming in caching frameworks has been widely investigated to minimize the normalized delivery time [20]–[23].

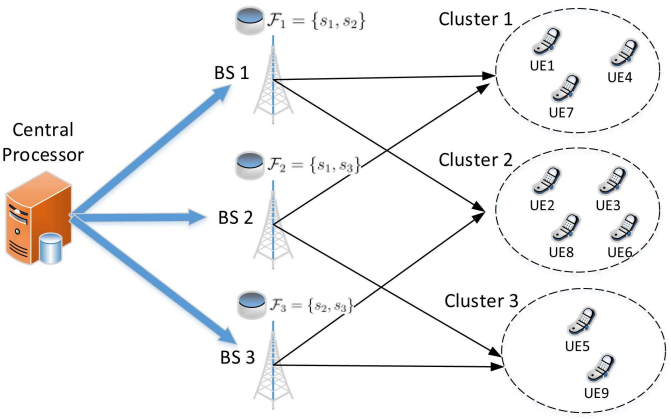


Fig. 1. An example of content-centric communications

where $\delta > 0$ is the skew parameter and $\sum_{f=1}^{N_c} p_f = 1$. The popularity distribution becomes more skewed towards the most popular content as the skew parameter δ increases. In contrast, less popular content items are more likely to be requested whenever δ is low. UEs request content items with the probability according to their popularity. Let $\mathcal{M} = \{s_1, \ldots s_M\}$, with $1 \leq M \leq N_u$ be the set of the requested contents. A UE that requests s_m that is missed in any SBS cache, acquires the service of the N_f nearest SBSs. Accordingly, the CP must fetch such cache-missed s_m to these SBSs.

Define $\mathcal{K}_m$ as the cluster of those UEs who request the same content item s_m

$$\mathcal{K}_m \triangleq \{1_m, \ldots, |\mathcal{K}_m|_m\}, |\mathcal{K}_m| \leq N_u.$$

Accordingly,

$$\mathcal{Q}_m \triangleq \{m_1, \ldots, m_{|\mathcal{Q}_m|}\} \subset \mathcal{Q}, |\mathcal{Q}_m| \leq N_{BS}$$

is the set of SBSs that have content item s_m in their disposal. The UEs are clustered according to their requested content items, but not relying on their locations, so $\mathcal{Q}_m \cap \mathcal{Q}_n$ may be not empty.

Each s_m is beamformed by vector $\mathbf{w}_{m_i} \in \mathbb{C}^{N_t}$ at SBS $m_i \in \mathcal{Q}_m$ before the transmission. The signal received at UE k_m is

$$y_{k_m} = \sum_{i=1}^{|\mathcal{Q}_m|} \mathbf{h}_{k_m,m_i}^H \mathbf{w}_{m_i} s_m + \sum_{\ell \neq m} \sum_{j=1}^{|\mathcal{Q}_\ell|} \mathbf{h}_{k_m,\ell_j}^H \mathbf{w}_{\ell_j} s_\ell + n_{k_m}, \tag{1}$$

where $\mathbf{h}_{k_m,\ell_j} \in \mathbb{C}^{N_t}$ is the channel vector from SBS ℓ_j to UE k_m and $n_{k_m} \sim \mathcal{CN}(\mu, \sigma_n^2)$ is the background noise. For

$$\mathbf{w}_m \triangleq \begin{bmatrix} \mathbf{w}_{m_1} \\ \ldots \\ \mathbf{w}_{m_{|\mathcal{Q}_m|}} \end{bmatrix} \in \mathbb{C}^{N_t.|\mathcal{Q}_m|}, \mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_m),$$

the signal-to-interference-plus-noise-ratio (SINR) at UE k_m is

$$\gamma_{k_m}(\mathbf{w}) = \frac{|\mathbf{h}_{k_m,m}^H \mathbf{w}_m|^2}{\sum_{\ell \neq m} |\mathbf{h}_{k_m,\ell}^H \mathbf{w}_\ell|^2 + \sigma^2}. \tag{2}$$

The throughput of s_m is defined by

$$\rho_m(\mathbf{w}) \triangleq \min_{k=1,\ldots,|\mathcal{K}_m|} \ln(1+\gamma_{k_m}(\mathbf{w})),$$

so QoS for multicasting s_m is set as

$$\rho_m(\mathbf{w}) \geq \gamma_m, m = 1,\ldots,M, \tag{3}$$

where γ_m is the delivery rate threshold for s_m.

Given a power budget P_{max}, the power constraint at each SBS $n \in \mathcal{Q}$ is

$$\sum_{m:\ n\in Q_m} \|\mathbf{w}_{m_n}\|^2 \leq P_{max}, n = 1,\ldots,N_{BS}. \tag{4}$$

The total power consumption is defined as

$$\tau(\mathbf{w}) = \alpha \sum_{n=1}^{N_{BS}} \sum_{m:\ n\in Q_m} \|\mathbf{w}_{m_n}\|^2 + P_{non}, \tag{5}$$

where $\alpha > 1$ is the reciprocal of the drain efficiency of the amplifier of SBSs and

$$P_{non} = N_t N_{BS} P_a,$$

with the per-antenna circuit power P_a of SBSs.

For a cache-missed s_m, the backhaul cost in fetching to a SBS via the backhaul link is its delivery rate. Thus, the total cost for fetching s_m to SBSs in cluster $\mathcal{Q}_m$ is [2]

$$|\mathcal{Q}_m|\rho_m(\mathbf{w}). \tag{6}$$

Let $\mathcal{M}_F$ be the set of the cache-missed contents. The total backhaul cost is

$$\sum_{m:s_m\in\mathcal{M}_F} |\mathcal{Q}_m|\rho_m(\mathbf{w}). \tag{7}$$

To have a computationally tractable formulation, instead of (7) we use its upper bound

$$\chi(\mathbf{w}) \triangleq \sum_{m:s_m\in\mathcal{M}_F} \frac{|\mathcal{Q}_m|}{|\mathcal{K}_m|} \sum_{k=1}^{|\mathcal{K}_m|} \ln(1+\gamma_{k_m}(\mathbf{w})). \tag{8}$$

In this paper we consider the following optimization problem

$$\max_{\mathbf{w}\triangleq(\mathbf{w}_1,\ldots,\mathbf{w}_M)} \frac{\sum_{m=1}^M \rho_m(\mathbf{w})}{\tau(\mathbf{w})+\chi(\mathbf{w})} \quad \text{s.t.} \quad (3),(4). \tag{9}$$

The objective function in (9) expresses the network cost efficiency because its numerator is the sum content throughput while its denominator is the sum of the consumption power and backhaul cost for fetching the cache-missed content items. As such, (9) is the problem of maximizing the cost efficiency subject to the QoS constraint (3) and power constraint (4). When there is no backhaul cost term $\chi(\mathbf{w})$ in the denominator, the objective function in (9) is the conventional EE and (9) is the problem of EE maximization.

For problem (9), like [17], the conventional assumption is that full channel state information (CSI) is available, which is practical for UDNs, where the propagation environment does not change rapidly due to the slow mobility of their UEs, making channel estimation effective and not costly.

[2]Recall that $|\mathcal{K}_m|$ is the number of SBSs that transmit s_m

One can see that (9) is a large-dimensional nonconvex optimization problem, which is computationally prohibitive in general. Our next sections are devoted to seeking the beamformers $\mathbf{w}_{m_n}$ in the class of ZF or GZF beamforming, under which problem (9) is transformed to that of their power allocation with the problem dimension essentially reduced, paving effective computation.

III. Zero-forcing beamforming

A. Specialized optimization formulation

Under the definitions

$$\mathbf{y}_m \triangleq \begin{bmatrix} y_{1_m} \\ \ldots \\ y_{|\mathcal{K}_m|_m} \end{bmatrix} \in \mathbb{C}^{|\mathcal{K}_m|}, \mathbf{h}_{k_m,\ell} \triangleq \begin{bmatrix} \mathbf{h}_{k_m,\ell_1} \\ \ldots \\ \mathbf{h}_{k_m,\ell_{|\mathcal{Q}_\ell|}} \end{bmatrix} \in \mathbb{C}^{N_t.|\mathcal{Q}_\ell|},$$

$$\mathbf{n}_m \triangleq \begin{bmatrix} n_{1_m} \\ \ldots \\ n_{|\mathcal{K}_m|_m} \end{bmatrix} \in \mathbb{C}^{|\mathcal{K}_m|},$$

and

$$\mathbf{H}_{m,\ell} \triangleq \begin{bmatrix} \mathbf{h}_{1_m,\ell}^H \\ \ldots \\ \mathbf{h}_{|\mathcal{K}_m|_m,\ell}^H \end{bmatrix} \in \mathbb{C}^{|\mathcal{K}_m|\times(N_t.|\mathcal{Q}_\ell|)},$$

$$\mathbf{H}_m \triangleq \begin{bmatrix} \mathbf{H}_{1,m} \\ \ldots \\ \mathbf{H}_{M,m} \end{bmatrix} \in \mathbb{C}^{N_u\times(N_t.|\mathcal{Q}_m|)}$$

we can rewrite (1) in a compact form as

$$\begin{aligned} \begin{bmatrix} \mathbf{y}_1 \\ \ldots \\ \mathbf{y}_M \end{bmatrix} &= \begin{bmatrix} \mathbf{H}_{1,1} & \ldots & \mathbf{H}_{1,M} \\ \ldots & \ldots & \ldots \\ \mathbf{H}_{M,1} & \ldots & \mathbf{H}_{M,M} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 s_1 \\ \ldots \\ \mathbf{w}_M s_M \end{bmatrix} + \begin{bmatrix} \mathbf{n}_1 \\ \ldots \\ \mathbf{n}_M \end{bmatrix} \\ &= \sum_{m=1}^M \mathbf{H}_m \mathbf{w}_m s_m + \begin{bmatrix} \mathbf{n}_1 \\ \ldots \\ \mathbf{n}_M \end{bmatrix}. \end{aligned} \tag{10}$$

For $m = 1,\ldots,M$, define the interfering matrix

$$\mathbf{H}_{-m} = \begin{bmatrix} \mathbf{H}_{1,m} \\ \ldots \\ \mathbf{H}_{m-1,m} \\ \mathbf{H}_{m+1,m} \\ \ldots \\ \mathbf{H}_{M,m} \end{bmatrix} \in \mathbb{C}^{(N_u-|\mathcal{K}_m|)\times(N_t|\mathcal{Q}_m|)}. \tag{11}$$

Under the assumption

$$N_u - |\mathcal{K}_m| < N_t|\mathcal{Q}_m|, m = 1,\ldots,M, \tag{12}$$

that requires that each content s_m is available in a sufficient number of SBSs, we seek $\mathbf{w}_m$ in the ZF class

$$\mathbf{H}_{-m}\mathbf{w}_m = 0_{N_u-|\mathcal{K}_m|}, m = 1,\ldots,M, \tag{13}$$

which forces the multi-content interference to zero, or equivalently,

$$\mathbf{w}_m = \mathcal{N}_m \boldsymbol{x}_m, m = 1,\ldots,M, \tag{14}$$

where $\mathcal{N}_m \in \mathbb{C}^{(N_t.|\mathcal{Q}_m|)\times(N_t.|\mathcal{Q}_m|-\eta_m)}$ is the zero base of $\mathbf{H}_{-m}$, η_m is the rank of $\mathbf{H}_{-m}$ and $\boldsymbol{x}_m \in \mathbb{C}^{N_t.|\mathcal{Q}_m|-\eta_m}$. By partitioning

$$\mathcal{N}_m = \begin{bmatrix} \mathcal{N}_{m_1} \\ \cdots \\ \mathcal{N}_{m_{|\mathcal{Q}_m|}} \end{bmatrix}, \mathcal{N}_{m_i} \in \mathbb{C}^{N_t \times (N_t.|\mathcal{Q}_m|-\eta_m)},$$

it follows that

$$\mathbf{w}_{m_i} = \mathcal{N}_{m_i}\boldsymbol{x}_m, i = 1, \ldots, |\mathcal{Q}_m|. \tag{15}$$

The power constraint (4) at each SBS $n \in \mathcal{Q}$ becomes

$$\sum_{m:n\in Q_m} ||\mathcal{N}_{m_n}\boldsymbol{x}_m||^2 \leq P_{max}, n = 1, \ldots, N_{BS}. \tag{16}$$

The power consumption defined from (5) is the convex quadratic function

$$t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \alpha \sum_{n=1}^{N_{BS}} \sum_{m:\ n\in Q_m} ||\mathcal{N}_{m_n}\boldsymbol{x}_m||^2 + P_{non}.$$

The SINR in (2) at UE K_m is simplified to

$$\gamma_{k_m}(\boldsymbol{x}_m) = \frac{|\mathbf{h}_{k_m,m}^H \mathcal{N}_m \boldsymbol{x}_m|^2}{\sigma^2}, \tag{17}$$

while QoS constraint (3) becomes

$$\begin{aligned} & r_m(\boldsymbol{x}_m) \geq \gamma_m,\ m = 1, \ldots, M \\ \Leftrightarrow\quad & \Gamma(\boldsymbol{x}) \geq 1 \end{aligned} \tag{18}$$

for

$$r_m(\boldsymbol{x}_m) \triangleq \min_{k=1,\ldots,|\mathcal{K}_m|} \ln(1 + \frac{|\mathbf{h}_{k_m,m}^H \mathcal{N}_m \boldsymbol{x}_m|^2}{\sigma^2}),$$

and

$$\Gamma(\boldsymbol{x}) \triangleq \min_{m=1,\ldots,M} \min_{k=1,\ldots,|\mathcal{K}_m|} \frac{|\mathbf{h}_{k_m,m}^H \mathcal{N}_m \boldsymbol{x}_m|^2}{(e^{\gamma_m}-1)\sigma^2}.$$

The problem (9) under the ZF class (13) is specialized to the following optimization problem

$$\begin{aligned} \max_{\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\ \boldsymbol{x}_m)} & \mathcal{E}(\boldsymbol{x}) \triangleq \frac{\sum_{m=1}^M r_m(\boldsymbol{x}_m)}{t(\boldsymbol{x}) + \sum_{m:\ s_m\in\mathcal{M}_F} \frac{|\mathcal{Q}_m|}{|\mathcal{K}_m|}\bar{r}_m(\boldsymbol{x}_m)} \\ \text{s.t.} & \quad (16), (18), \end{aligned} \tag{19}$$

where

$$\bar{r}_m(\boldsymbol{x}_m) \triangleq \sum_{k=1}^{|\mathcal{K}_m|} \ln\left(1 + \frac{|\mathbf{h}_{k_m,m}^H \mathcal{N}_m \boldsymbol{x}_m|^2}{\sigma^2}\right).$$

To the authors' best knowledge, problem (19) has not been considered in the literature so far. The numerator of the objective function in (19) is not only nonconcave (preventing of exploiting Dinkelbach's iterations) but is non-differentiable (preventing of exploiting gradient-based methods such as Frank-and-Wolfe's), while its denominator is also a nonconvex function. The next subsection is devoted to its computation.

B. ZF Path-following Method

A path-following method starts from an initial feasible point $\boldsymbol{x}^{(0)} = (\boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_M^{(0)})$ for (19) and then at the κ-th iteration for $\kappa = 0, 1, \ldots$, it generates a feasible point $\boldsymbol{x}^{(\kappa+1)} = (\boldsymbol{x}_1^{(\kappa+1)}, \ldots, \boldsymbol{x}_M^{(\kappa+1)})$ which is better than the incumbent $\boldsymbol{x}^{(\kappa)} = (\boldsymbol{x}_1^{(\kappa)}, \ldots, \boldsymbol{x}_M^{(\kappa)})$.

1) Initialization step: To find a feasible point for (19) consider the problem

$$\max_{\boldsymbol{x}} \ \Gamma(\boldsymbol{x}) \quad \text{s.t.} \quad (16). \tag{20}$$

For easy of presentation, define the linear mapping

$$\lambda_{k_m,\ell}(\boldsymbol{x}_\ell) = \mathbf{h}_{k_m,\ell}^H \mathcal{N}_\ell \boldsymbol{x}_\ell. \tag{21}$$

For all $\boldsymbol{x}_m$ and $\boldsymbol{x}_m^{(\kappa)}$, it is true that

$$\begin{aligned} |\mathbf{h}_{k_m,m}^H \mathcal{N}_m \boldsymbol{x}_m|^2 &= |\lambda_{k_m,m}(\boldsymbol{x}_m)|^2 \\ &\geq \ell_{k_m,m}^{(\kappa)}(\boldsymbol{x}_m) \\ &\triangleq 2\Re\{\lambda_{k_m,m}^*(\boldsymbol{x}_m^{(\kappa)})\lambda_{k_m,m}(\boldsymbol{x}_m)\} \\ &\quad -|\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2, \end{aligned} \tag{22}$$

where the function $\ell_{k_m,m}^{(\kappa)}$ is affine. Initialized by any feasible point $\boldsymbol{x}_m^{(0)}$ for the convex constraint (16), at the κ-th iteration we solve the convex optimization problem[3]

$$\begin{aligned} \max_{\boldsymbol{x}} \ & \Lambda^{(\kappa)}(\boldsymbol{x}) \triangleq \min_{m=1,\ldots,M} \min_{k=1,\ldots,|\mathcal{K}_m|} \frac{\ell_{k_m,m}^{(\kappa)}(\boldsymbol{x}_m)}{(e^{\gamma_m}-1)\sigma^2} \\ \text{s.t.} \quad & (16), \end{aligned} \tag{23}$$

to generate $\boldsymbol{x}^{(\kappa+1)}$. Note that

$$\Lambda^{(\kappa)}(\boldsymbol{x}) \leq \Gamma(\boldsymbol{x}) \quad \forall\ \boldsymbol{x} \tag{24}$$

thanks to (22) and $\Lambda^{(\kappa)}(\boldsymbol{x}^{(\kappa)}) = \Gamma(\boldsymbol{x}^{(\kappa)})$, which can be easily verified. Moreover, $\Lambda^{(\kappa)}(\boldsymbol{x}^{(\kappa+1)}) > \Lambda^{(\kappa)}(\boldsymbol{x}^{(\kappa)})$ as far as $\boldsymbol{x}^{(\kappa+1)} \neq \boldsymbol{x}^{(\kappa)}$ because $\boldsymbol{x}^{(\kappa+1)}$ is the optimal solution of (23) while $\boldsymbol{x}^{(\kappa)}$ is its feasible point. Therefore,

$$\Gamma(\boldsymbol{x}^{(\kappa+1)}) \geq \Lambda^{(\kappa)}(\boldsymbol{x}^{(\kappa+1)}) > \Lambda^{(\kappa)}(\boldsymbol{x}^{(\kappa)}) = \Gamma(\boldsymbol{x}^{(\kappa)}).$$

Till reaching

$$\Gamma(\boldsymbol{x}^{(\kappa+1)}) \geq 1, \tag{25}$$

to satisfy the QoS constraint (18), we reset $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{x}^{(\kappa+1)}$ for the next iteration.

2) The κ-th iteration: Let $\mathbf{x}^{(\kappa)}$ be the feasible point for (19), which is found from the $(\kappa-1)$th iteration. By inequality (A.4) in the appendix,

$$\ln\left(1 + |\lambda_{k_m,m}(\boldsymbol{x}_m)|^2/\sigma^2\right) \geq r_{ZF,k_m,m}^{(\kappa)}(\boldsymbol{x}_m), \tag{26}$$

where

$$r_{ZF,k_m,m}^{(\kappa)}(\boldsymbol{x}_m) \stackrel{\triangle}{=} a_{k_m,m}^{(\kappa)} + b_{k_m,m}^{(\kappa)}(1 - \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2}{\ell_{k_m,m}^{(\kappa)}(\boldsymbol{x}_m)}), \tag{27}$$

[3]Function $\Lambda^{(\kappa)}(\boldsymbol{x})$ is concave as pointwise minimization of affine functions [30]

Algorithm 1 Path-following algorithm for ZF beamforming

Initialize Iterate (23) for a feasible point $\boldsymbol{x}^{(0)}$. Set $\kappa := 0$.
repeat
Solve the quadratic program (31) for the optimal solution $\boldsymbol{x}^{(\kappa+1)}$.
Reset $\kappa \leftarrow \kappa + 1$.
until convergence

which is a concave function, and

$$\begin{aligned} 0 < a_{k_m,m}^{(\kappa)} &= \ln(1 + |\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2/\sigma^2), \\ 0 < b_{k_m,m}^{(\kappa)} &= \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2/\sigma^2}{1 + |\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2/\sigma^2}. \end{aligned} \tag{28}$$

The function

$$r_{ZF,m}^{(\kappa)}(\boldsymbol{x}_m) \triangleq \min_{k=1,\ldots,|\mathcal{K}_m|} r_{ZF,k_m,m}^{(\kappa)}(\boldsymbol{x}_m),$$

is concave [4] and satisfies

$$r_{ZF,m}^{(\kappa)}(\boldsymbol{x}_m) \le r_m(\boldsymbol{x}_m) \tag{29}$$

with $r_{ZF,m}^{(\kappa)}(\boldsymbol{x}_m^{(\kappa)}) = r_m(\boldsymbol{x}_m^{(\kappa)})$, $m = 1, \ldots, M$.

On the other hand, by using inequality (A.7) in the appendix we obtain

$$\begin{aligned} \bar{r}_m(\boldsymbol{x}_m) \le & \bar{r}_{ZF,m}^{up,(\kappa)}(\boldsymbol{x}_m) \\ \triangleq & \sum_{k=1}^{|\mathcal{K}_m|} \left[\ln\left(1 + \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2}{\sigma^2}\right) \right. \\ & + \left(1 + \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2}{\sigma^2}\right)^{-1} \\ & \left. \times \left(\frac{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}{\sigma^2} - \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m^{(\kappa)})|^2}{\sigma^2} \right) \right]. \end{aligned} \tag{30}$$

The function $\bar{r}_{ZF,m}^{up,(\kappa)}(\boldsymbol{x}_m)$ is convex and satisfies $\bar{r}_m(\boldsymbol{x}_m^{(\kappa)}) = \bar{r}_{ZF,m}^{up,(\kappa)}(\boldsymbol{x}_m^{(\kappa)})$.

At the κth iteration, we solve the following convex optimization problem to generalize the next feasible point $\boldsymbol{x}^{(\kappa+1)}$ for (19)

$$\begin{aligned} \max_{\boldsymbol{x}} \ \Upsilon(\boldsymbol{x}) \triangleq & \sum_{m=1}^{M} r_{ZF,m}^{(\kappa)}(\boldsymbol{x}_m) - \mathcal{E}(\boldsymbol{x}^{(\kappa)}) \\ & \times \left(t(\boldsymbol{x}) + \sum_{m:s_m \in \mathcal{M}_F} \frac{|\mathcal{Q}_m|}{|\mathcal{K}_m|} \bar{r}_{ZF,m}^{up,(\kappa)}(\boldsymbol{x}_m) \right) \\ \text{s.t.} \quad & (16), \Lambda^{(\kappa)}(\boldsymbol{x}) \ge 1. \end{aligned} \tag{31}$$

Due to (24), each feasible point for (31) is also feasible for (19). Note that $\boldsymbol{x}^{(\kappa)}$ is a feasible point for (31) with

$$\Upsilon(\boldsymbol{x}^{(\kappa)}) = 0.$$

[4]The function $r_{ZF,m}^{(\kappa)}$ is concave as a pointwise minimum of concave functions [30]

Therefore, as far as $\boldsymbol{x}^{(\kappa+1)} \neq \boldsymbol{x}^{(\kappa)}$, it is true that

$$\begin{aligned} & \Upsilon(\boldsymbol{x}^{(\kappa+1)}) > 0 \\ \Leftrightarrow \quad & \sum_{m=1}^{M} r_{ZF,m}^{(\kappa)}(\boldsymbol{x}_m^{(\kappa+1)}) / \left[t(\boldsymbol{x}^{(\kappa+1)}) \right. \\ & \left. + \textstyle\sum_{m:s_m \in \mathcal{M}_F} \frac{|\mathcal{Q}_m|}{|\mathcal{K}_m|} \bar{r}_{ZF,m}^{up,(\kappa)}(\boldsymbol{x}_m^{(\kappa+1)}) \right] > \mathcal{E}(\boldsymbol{x}^{(\kappa)}). \end{aligned}$$

But by (29) and (30),

$$\begin{aligned} \mathcal{E}(\boldsymbol{x}^{(\kappa+1)}) \ge & \sum_{m=1}^{M} r_{ZF,m}^{(\kappa)}(\boldsymbol{x}_m^{(\kappa+1)}) / \left[t(\boldsymbol{x}^{(\kappa+1)}) \right. \\ & \left. + \sum_{m:s_m \in \mathcal{M}_F} \frac{|\mathcal{Q}_m|}{|\mathcal{K}_m|} \bar{r}_m(\boldsymbol{x}_m^{(\kappa+1)}) \right], \end{aligned}$$

so

$$\mathcal{E}(\boldsymbol{x}^{(\kappa+1)}) > \mathcal{E}(\boldsymbol{x}^{(\kappa)}), \tag{32}$$

implying the $\boldsymbol{x}^{(\kappa+1)}$ is a better feasible point for (19) than $\boldsymbol{x}^{(\kappa)}$. As such, Algorithm 1 at least converges to a locally optimal solution of (19) [31].

IV. Generalized zero-forcing beamforming

A. Specialized optimization Formulation

Without the assumption (12), a nonzero $\mathbf{w}_m$ to satisfy the ZF condition (13) may not exist. A natural approach is to satisfy the zero-forcing condition for the rank-η_m best approximation of $\mathbf{H}_{-m}$ for $\eta_m < N_t.|\mathcal{Q}_m|$ as follows. Make the singular-value decomposition (SVD)

$$\mathbf{H}_{-m} = \mathbf{U}_{-m}\Sigma_{-m}\mathbf{V}_{-m},$$

with unitary matrices $\mathbf{U}_{-m} \in \mathbb{C}^{(N_u - |\mathcal{K}_m|)\times(N_u - |\mathcal{K}_m|)}$ and $\mathbf{V}_{-m} \in \mathbb{C}^{(N_t.|\mathcal{Q}_m|)\times(N_t.|\mathcal{Q}_m|)}$ and diagonal Σ_{-m} with singular values in decreasing order on its diagonal. By keeping only the η_m largest eigenvalues and resetting the other on diagonal to zero to obtain matrix $\tilde{\Sigma}_{-m}$, the best rank-η_m approximation of $\mathbf{H}_{-m}$ is obtained as

$$\tilde{\mathbf{H}}_{-m} = \mathbf{U}_{-m}\tilde{\Sigma}_{-m}\mathbf{V}_{-m}.$$

By the Eckart and Young Theorem,

$$||\mathbf{H}_{-m} - \tilde{\mathbf{H}}_{-m}||_F = \sigma_{\eta_m+1},$$

where σ_{η_m+1} is the (η_m+1)-th largest singular eigenvalue of $\mathbf{H}_{-m}$. For simplicity, we choose η_m as

$$\eta_m = \lfloor \frac{N_t.|\mathcal{Q}_m|}{2} \rfloor. \tag{33}$$

We thus seek $\mathbf{w}_m$ in the GZF class

$$\tilde{\mathbf{H}}_{-m}\mathbf{w}_m = 0, \tag{34}$$

or equivalently $\mathbf{w}_m$ and $\mathbf{w}_{m_i}$ are in forms (14) and (15), where $\mathcal{N}_m \in \mathbb{C}^{(N_t.|\mathcal{Q}_m|)\times(N_t.|\mathcal{Q}_m|-\eta_m)}$ is the zero base of $\tilde{\mathbf{H}}_{-m}$.

Define the convex quadratic function

$$\beta_{k_m,m}(\boldsymbol{x}) = \sum_{\ell \neq m} |\lambda_{k_m,\ell}(\boldsymbol{x}_\ell)|^2 + \sigma^2. \tag{35}$$

Recalling definition (21), SINR (2) at UE k_m is now specialized to

$$\gamma_{k_m}(\boldsymbol{x}) = \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}{\beta_{k_m,m}(\boldsymbol{x})}, \tag{36}$$

while the QoS constraint (3) is

$$\begin{aligned} & r_m(\boldsymbol{x}) \geq \gamma_m, \ m = 1, \ldots, M, \\ \Leftrightarrow \quad & \Gamma(\boldsymbol{x}) \geq 1 \\ \Leftrightarrow \quad & \Psi(\boldsymbol{x}) \leq 1 \end{aligned} \tag{37}$$

for

$$r_m(\boldsymbol{x}) \triangleq \min_{k=1,\ldots,|\mathcal{K}_m|} \ln(1 + \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}{\beta_{k_m,m}(\boldsymbol{x})}),$$

and

$$\Gamma(\boldsymbol{x}) \triangleq \min_{m=1,\ldots,M} \min_{k=1,\ldots,|\mathcal{K}_m|} \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}{(e^{\gamma_m}-1)\beta_{k_m,m}(\boldsymbol{x})},$$

and

$$\begin{aligned} \Psi(\boldsymbol{x}) \quad &\triangleq \quad [\Gamma(\boldsymbol{x})]^{-1} \\ &= \quad \max_{m=1,\ldots,M} \max_{k=1,\ldots,|\mathcal{K}_m|} \frac{(e^{\gamma_m}-1)\beta_{k_m,m}(\boldsymbol{x})}{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}. \end{aligned}$$

It should be realized that

$$\lambda_{k_m,l}(\boldsymbol{x}_l) = 0 \tag{38}$$

whenever $\text{rank}(\mathbf{H}_{-m}) = \eta_m$, and

$$|\lambda_{k_m,l}(\boldsymbol{x}_l)|^2 \leq \sigma^2_{\eta_m+1} ||\mathcal{N}_\ell \boldsymbol{x}_\ell||^2. \tag{39}$$

For the convex quadratic function $t(\boldsymbol{x})$ defined by (5), the problem (9) under the GZF condition (34) is specialized to the following optimization problem:

$$\begin{aligned} \max_{\boldsymbol{x}} \ & \mathcal{E}(\boldsymbol{x}) \triangleq \sum_{m=1}^{M} r_m(\boldsymbol{x}) / \left[t(\boldsymbol{x}) + \sum_{m:s_m \in \mathcal{M}_F} \frac{|\mathcal{Q}_m|}{|\mathcal{K}_m|} \bar{r}_m(\boldsymbol{x}) \right] \\ \text{s.t.} \quad & (16), (37), \end{aligned} \tag{40}$$

where

$$\bar{r}_m(\boldsymbol{x}) \triangleq \sum_{k=1}^{|\mathcal{K}_m|} \ln\left(1 + \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}{\beta_{k_m,m}(\boldsymbol{x})}\right).$$

B. GZF Path-following Method

1) Initialization step: To find a feasible point for (40) we need to consider the problem

$$\min_{\boldsymbol{x}} \ \Psi(\boldsymbol{x}) \quad \text{s.t.} \quad (16). \tag{41}$$

Recalling definition (21) and (22) for the linear functions $\lambda_{k_m,m}(\boldsymbol{x}_m)$ and $\ell^{(\kappa)}_{k_m,m}(\boldsymbol{x}_m)$, initialized by any feasible point $\boldsymbol{x}^{(0)}$ for the convex constraint (16), at the κ-th iteration we solve the convex optimization problem

$$\min_{\boldsymbol{x}} \ \Psi^{(\kappa)}(\boldsymbol{x}) \quad \text{s.t.} \quad (16), \tag{42a}$$

$$\ell^{(\kappa)}_{k_m,m}(\boldsymbol{x}_m) > 0, k = 1, \ldots, |\mathcal{K}_m|; m = 1, \ldots, M, \tag{42b}$$

for[5]

$$\Psi^{(\kappa)}(\boldsymbol{x}) \triangleq \max_{m=1,\ldots,M} \max_{k=1,\ldots,|\mathcal{K}_m|} \frac{(e^{\gamma_m}-1)\beta_{k_m,m}(\boldsymbol{x})}{\ell^{(\kappa)}_{k_m,m}(\boldsymbol{x}_m)}$$

to generate $\boldsymbol{x}^{(\kappa+1)}_m$. Note that $\Psi^{(\kappa)}(\boldsymbol{x}) \geq \Psi(\boldsymbol{x})$ by (22), so

$$\Psi(\boldsymbol{x}^{(\kappa+1)}) \leq \Psi^{(\kappa)}(\boldsymbol{x}^{(\kappa+1)}) < \Psi^{(\kappa)}(\boldsymbol{x}^{(\kappa)}) = \Psi(\boldsymbol{x}^{(\kappa)}).$$

Till reaching

$$\Psi(\boldsymbol{x}^{(\kappa+1)}) \leq 1 \tag{43}$$

to satisfy the QoS constraint (37), we then reset $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{x}^{(\kappa+1)}$ for the next iterative process.

2) The κ-th iteration: Let $\mathbf{x}^{(\kappa)}$ be the feasible point for (40) found from the $(\kappa-1)$th iteration. By inequality (A.2) in the appendix,

$$\ln(1 + \frac{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}{\beta_{k_m,m}(\boldsymbol{x})}) \geq r^{(\kappa)}_{GZF,k_m,m}(\boldsymbol{x}), \tag{44}$$

over the trust region (42b), where

$$\begin{aligned} r^{(\kappa)}_{GZF,k_m,m}(\boldsymbol{x}) \triangleq & a^{(\kappa)}_{k_m,m} + b^{(\kappa)}_{k_m,m} \\ & \times (2 - \frac{|\lambda_{k_m,m}(\boldsymbol{x}^{(\kappa)}_m)|^2}{\ell^{(\kappa)}_{k_m,m}(\boldsymbol{x}_m)} - \frac{\beta_{k_m,m}(\boldsymbol{x})}{\beta_{k_m,m}(\boldsymbol{x}^{(\kappa)})}), \end{aligned} \tag{45}$$

which is a concave function with

$$\begin{aligned} 0 < a^{(\kappa)}_{k_m,m} &= \ln(1 + |\lambda_{k_m,m}(\boldsymbol{x}^{(\kappa)}_m)|^2 / \beta_{k_m,m}(\boldsymbol{x}^{(\kappa)})), \\ 0 < b^{(\kappa)}_{k_m,m} &= \frac{|\lambda_{k_m,m}(\boldsymbol{x}^{(\kappa)}_m)|^2 / \beta_{k_m,m}(\boldsymbol{x}^{(\kappa)})}{1 + |\lambda_{k_m,m}(\boldsymbol{x}^{(\kappa)}_m)|^2 / \beta_{k_m,m}(\boldsymbol{x}^{(\kappa)})}. \end{aligned} \tag{46}$$

The function

$$r^{(\kappa)}_{GZF,m}(\boldsymbol{x}) \triangleq \min_{k_m=1,\ldots,|\mathcal{K}_m|} r^{(\kappa)}_{GZF,k_m,m}(\boldsymbol{x})$$

is concave, which satisfies $r^{(\kappa)}_{GZF,m}(\boldsymbol{x}) \leq r_m(\boldsymbol{x})$ and $r^{(\kappa)}_{GZF,m}(\boldsymbol{x}^{(\kappa)}) = r_m(\boldsymbol{x}^{(\kappa)})$.

On the other hand, by using the inequality (A.5) in the appendix and also (44), we obtain

$$\begin{aligned} \bar{r}_m(\boldsymbol{x}) \quad &\leq \quad \bar{r}^{up,(\kappa)}_{GZF,m}(\boldsymbol{x}) \\ &\triangleq \quad \frac{1}{|\mathcal{K}_m|} \sum_{k=1}^{|\mathcal{K}_m|} \left[\ln\left(1 + \frac{|\lambda_{k_m,m}(\boldsymbol{x}^{(\kappa)}_m)|^2}{\beta_{k_m,m}(\boldsymbol{x}^{(\kappa)})}\right) \right. \\ & \quad + \left(1 + \frac{|\lambda_{k_m,m}(\boldsymbol{x}^{(\kappa)}_m)|^2}{\beta_{k_m,m}(\boldsymbol{x}^{(\kappa)})}\right)^{-1} \\ & \quad \left. \times \left(\frac{|\lambda_{k_m,m}(\boldsymbol{x}_m)|^2}{\mathcal{L}^{(\kappa)}_{k_m}(\boldsymbol{x}) + \sigma^2} - \frac{|\lambda_{k_m,m}(\boldsymbol{x}^{(\kappa)}_m)|^2}{\beta_{k_m,m}(\boldsymbol{x}^{(\kappa)})} \right) \right], \end{aligned}$$

over the trust region

$$\mathcal{L}^{(\kappa)}_{k_m}(\boldsymbol{x}) > 0, k = 1, \ldots, |\mathcal{K}_m|, \tag{47}$$

for

$$\begin{aligned} \mathcal{L}^{(\kappa)}_{k_m}(\boldsymbol{x}) \quad &\triangleq \quad \sum_{\ell \neq m} \left[2\Re\{(\lambda_{k_m,\ell}(\boldsymbol{x}^{(\kappa)}_\ell))^* \lambda_{k_m,\ell}(\boldsymbol{x}_\ell)\} \right. \\ & \quad \left. - |\lambda_{k_m,\ell}(\boldsymbol{x}^{(\kappa)}_\ell)|^2 \right]. \end{aligned}$$

[5]Function $\Psi^{(\kappa)}(\boldsymbol{x})$ is convex as pointwise maximum of convex functions [30]

Algorithm 2 Path-following algorithm for GZF beamforming

Initialize Iterate (42) for a feasible point $\boldsymbol{x}^{(0)}$. Set $\kappa := 0$.
repeat
Solve the convex optimization problem (48) for the optimal solution $\boldsymbol{x}^{(\kappa+1)}$.
Reset $\kappa \leftarrow \kappa + 1$.
until convergence

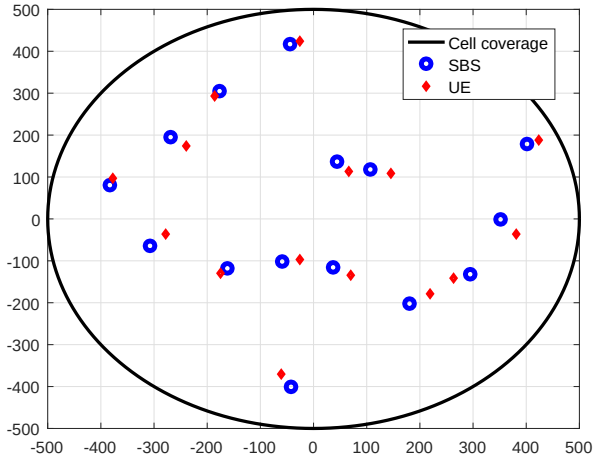


Fig. 2. Simulation scenario with $N_{BS} = 15$ SBSs and $N_u = 15$ UEs uniformly distributed in the cell.

The function $\bar{r}^{up,(\kappa)}_{GZF,m}(\boldsymbol{x})$ is convex and satisfies $\bar{r}_m(\boldsymbol{x}^{(\kappa)}) = \bar{r}^{up,(\kappa)}_{GZF,m}(\boldsymbol{x}^{(\kappa)})$. We solve the following convex optimization problem to generate $\boldsymbol{x}^{(\kappa+1)}$

$$\max_{\boldsymbol{x}} \sum_{m=1}^{M} r^{(\kappa)}_{GZF,m}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{x}^{(\kappa)}) \times \left(t(\boldsymbol{x}) + \sum_{m:s_m \in \mathcal{M}_F} \frac{|\mathcal{Q}_m|}{|\mathcal{K}_m|} \bar{r}^{up,(\kappa)}_{GZF,m}(\boldsymbol{x}) \right)$$
$$\text{s.t.} \quad (16), (42b), (47), \Psi^{(\kappa)}(\boldsymbol{x}) \leq 1. \tag{48}$$

Similarly, we can easily prove (32), and like Algorithm 1, Algorithm 2 at least converges to a locally optimal solution of the problem (40).

V. Simulation Results

Consider an UDN scenario as illustrated by Fig. 2. There are 15 SBSs ($N_{BS} = 15$), each is equipped with four antennas ($N_t = 4$), and 15 single-antenna UEs ($N_u = 15$), which are uniformly distributed in the cell. The channel vector $\mathbf{h}_{k_m,m_i}$ between the SBS $m_i \in \mathcal{Q}_m$ and UE $k_m \in \mathcal{K}_m$ is defined as $\mathbf{h}^H_{k_m,m_i} = \sqrt{10^{-\sigma_{pl}/10}}\tilde{\mathbf{h}}_{k_m,m_i}$, where the path loss component σ_{pl} is modeled as

$$\sigma_{pl} = 148.1 + 37.6 \log_{10}(d_{k_m,m_i}), \tag{49}$$

with the distance d_{k_m,m_i} in kilometers, and $\tilde{\mathbf{h}}_{k_m,m_i}$ determines small-scale effects. The other parameters in Table I are taken from [32] and [33].

TABLE I
Parameter Settings

Parameter	Value
Radius of cell	500 m
Coverage of SBS	40 m
Carrier frequency/ Bandwidth(Bw)	2 GHz / 10 MHz
Shadowing standard deviation	8 dB
Efficiency of power amplifiers ($1/\alpha$)	0.052
Noise power density	-174 dBm/Hz
The circuit power per antenna	5.6*1e-3 W

There are total 30 content items ($N_c = 30$). Except in Figs. 6 and 7, each SBS's cache is assumed to store 20 content items ($N_s = 20$). Set $\gamma_m = \gamma_{th}, m = 1, \ldots, M$, for simplicity. Let us consider the three following caching strategies.

- *Most Popular Caching:* The most popular content items are cached by each of SBSs until its cache is full. The cached content items thus are the same for all SBSs, so $|\mathcal{F}_q| \equiv N_s$ and $|\mathcal{Q}_m| \equiv N_{BS}$. The UEs' requests are fully matched with the most popular content items whenever δ is large. The cache hit ratio is low and thus the backhaul links are costly whenever δ is small because the less popular content items are likely requested.
- *Fair Caching:* All content items are equally distributed in the system to boost the cache-hit ratio. Namely, each content item is stored by random $\lfloor \frac{N_{BS}N_s}{N_c} \rfloor$ SBSs. It follows that $|\mathcal{F}_q| \leq N_s$ and $|\mathcal{Q}_m| \equiv \lfloor \frac{N_{BS}N_s}{N_c} \rfloor$. The cache-hit probability is one, so the backhaul link cost is zero.
- *Probabilistic Caching:* each SBS decides on its cached content items based on the contents' popularity, i.e. it selects a content item with the probability equal to the content popularity. More popular content items have a higher potential to be cached at SBSs. It follows that $|\mathcal{F}_q| \leq N_s$ and $|\mathcal{Q}_1| > \cdots > |\mathcal{Q}_{N_c}|$.

A. The effects of QoS requirements

To investigate the effect of QoS requirements on the cost efficiency, we fix the skew parameter δ at the low value 0.2 so the UEs will request both popular and less popular content items, and the power budget P_{max} at 1 W. We employ the Most Popular Caching strategy to make sure that there are cache-missed content items and thus the backhaul cost cannot be ignored. Algorithm 2 is implemented since the ZF condition (12) is not fulfilled.

It can be observed from Fig. 3 that the cost efficiency at lower QoS is better than that at higher QoS. The increase of the rate thresholds gives rise to failure in providing the required QoS. A higher QoS also costs more backhaul links. The SBSs must use more transmit power to deliver the contents, which also downgrades the cost efficiency. The cost efficiency is highest by fetching each cache-missed content item to a single SBS, i.e. $N_f = 1$. For $N_f = 3$ and $N_f = 5$, the cost efficiency is lower due to the increased backhaul cost.

B. The effects of transmit power

The skew parameter δ is still fixed at 0.2 but the Fair Caching strategy is employed to avoid the backhaul cost.

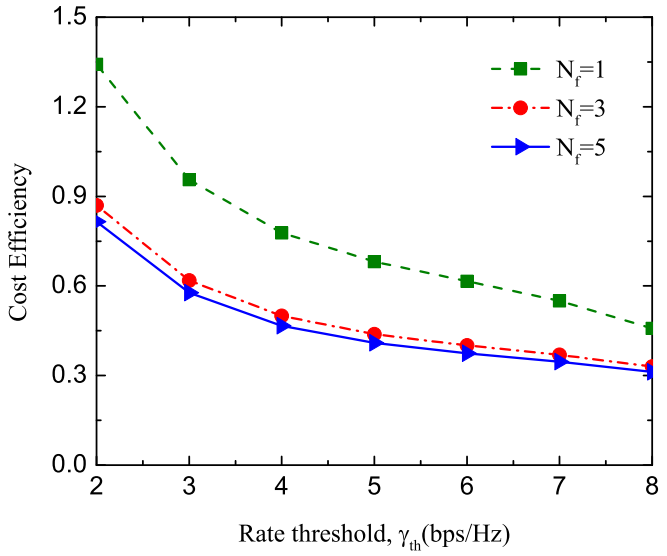


Fig. 3. The cost efficiency versus minimum rate threshold γ_{th}.

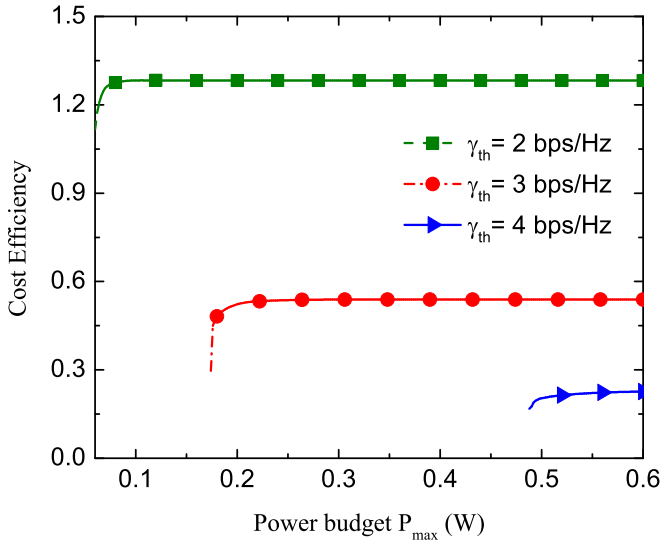


Fig. 4. The cost efficiency versus power budget P_{max}.

Algorithm 1 is implemented since the ZF condition (12) is fulfilled. Fig. 4 plots the cost efficiency versus the power budget under different QoSs. The minimal power budget $P_{\max}$ for providing the QoS of $\gamma_{th} = 2$ bps/Hz, $\gamma_{th} = 3$ bps/Hz, and $\gamma_{th} = 4$ bps/Hz is $= 0.06$ W, 0.18 W, and 0.48 W, respectively. Increasing of transmit power exploits more collaborative transmission. As expected, the cost efficiency increases and then quickly saturates when the sum throughput cannot be improved by using more transmit power.

C. The effects of content popularity and caching strategies

We fix $\gamma_{th} = 2$ bps/Hz, $P_{max} = 1$ W and $N_f = 1$ to investigate the effects of content popularity on the cost efficiency. For each value of the skew parameter δ, the cost efficiency is the average over running 100 simulation trials.

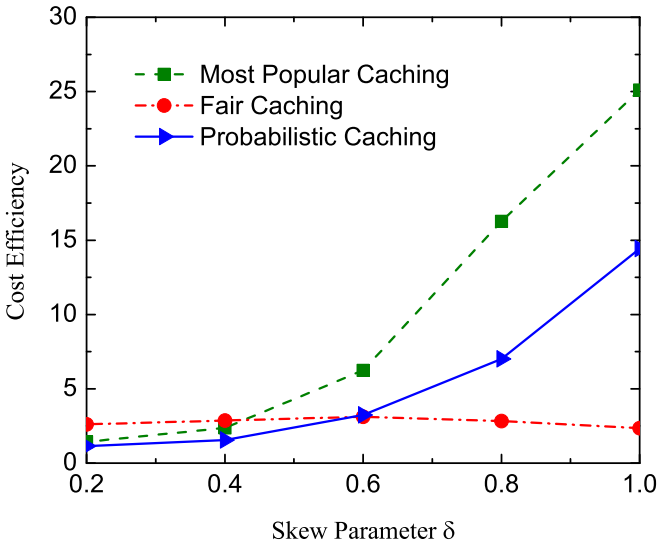


Fig. 5. The cost efficiency versus skew parameter δ.

Based on the satisfaction/dissatisfaction of the ZF condition (12), which is dependent on the caching strategy used and the value of the skew parameter α, Algorithm 1 or Algorithm 2 can be implemented. By Fig. 5, the Fair Caching strategy outperforms other two caching strategies at low δ when the backhaul cost under the former is zero but is sizable under the latter two. The Most Popular Caching strategy is advantageous with δ increased. As δ approaches 1, the most popular content items, which are cached at SBSs under the Most Popular Caching strategy have higher probability to be requested, making its cost efficiency significantly increased.

Although the Fair Caching strategy outperforms the Probabilistic Caching strategy for $\delta < 0.6$, the former retains low with δ increased, while the latter exploit the benefits from caching the popular content items.

D. The effects of caching capacity

Recalling that the SBS cache is assumed to store N_s content items, Figs. 6 and 7 plot the cost efficiency versus N_s, under $\gamma_{th} = 2$ bps/Hz, $P_{max} = 1$ W and $N_f = 1$, which is the average over running 100 simulation trials. Algorithm 1 is needed for those scenarios such that the ZF condition (12) is guaranteed. Otherwise, Algorithm 2 is needed. With the low value 0.2 of the skew parameter δ, Fig. 6 shows that the cost efficiency under the Fair Caching strategy dramatically increases when N_s approaches 24. The reason is that under this strategy, more SBSs are engaged in the collaborative delivery of each content item when N_s increases, which upgrade the total delivery throughput and thus improve the cost efficiency. The collaboration gain is so strongly reduced with $N_s < 12$ that QoS cannot be met. In addition, less popular content items are likely to be requested for $\delta = 0.2$. As a result, the cost efficiency under the Most Popular Caching strategy and the Probabilistic Caching strategy are mainly affected by the backhaul cost and thus exhibits an unpredictable behavior.

For the high value 1 of the skew parameter δ, the cost

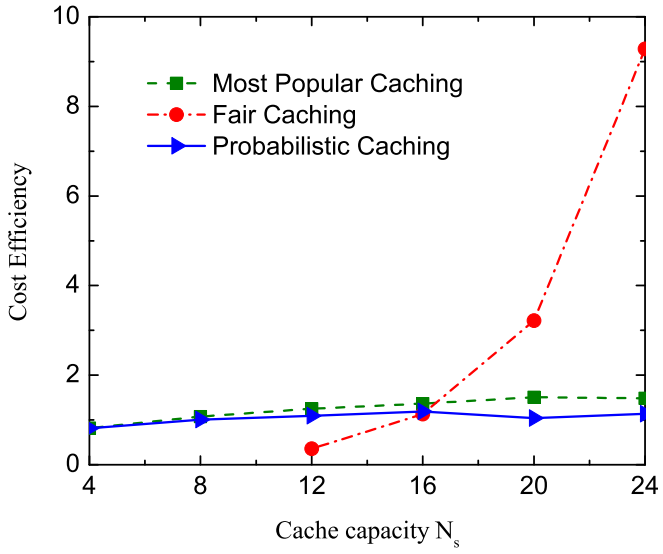


Fig. 6. The cost efficiency versus caching capacity N_s with $\delta = 0.2$.

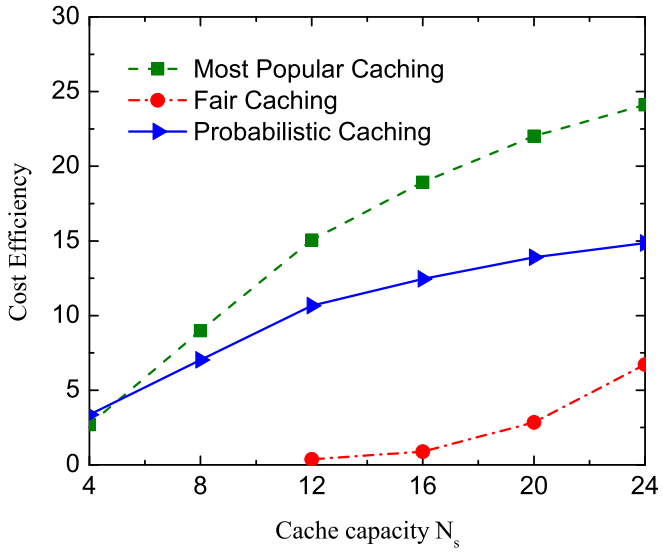


Fig. 7. The cost efficiency versus caching capacity N_s with $\delta = 1$.

efficiency under the Most Popular Caching strategy and the Probabilistic Caching strategy in Fig. 7 is high since the requested content items are well matched with the cached content items. In contrast, the cost efficiency under the Fair Caching strategy is remarkably low because its collaboration gain to deliver the requested content items is much smaller than that under the other two strategies. Also, the cache miss ratio is increased with the caching capacity decreased, leading to more backhaul cost and thus making the cost efficiency strongly reduced.

E. Convergence of the proposed algorithms

The typical convergence behavior of Algorithm 1 and Algorithm 2 is shown in Fig. 8. The settings are the same as those used in Fig. 4 with $P_{max} = 0.5$ W and Fig. 3 with $N_f = 5$. Both Algorithm 1 and Algorithm 2 converge to the optimal solution within 10 iterations.

VI. Conclusions

The paper has proposed a collaborative multicast beamforming to maximize the cost efficiency in terms of the ratio of the sum content throughput and the sum of the consumption power and backhaul cost for fetching cache-missed content items while guaranteeing their QoS. Furthermore, the ZF and GZF beamforming were also introduced to force the multi-content interference to zero or mitigate it. The computational low-complexity procedures, which invoke a simple convex optimization at each iteration, have been developed for computation. Numerical results have been provided to give insights into the impact of content-centric caching on the delivery performance.

Appendix: Fundamental Inequalities

For every $x > 0$, $y > 0$, $\bar{x} > 0$, and $\bar{y} > 0$ [34]

$$\ln(1 + 1/xy) \geq \ln(1 + 1/\bar{x}\bar{y}) + \frac{1/\bar{x}\bar{y}}{1 + 1/\bar{x}\bar{y}}(2 - x/\bar{x} - y/\bar{y}). \tag{A.1}$$

Particularly,

$$\begin{aligned}\ln(1 + |z|^2/y) \geq& \ln(1 + |\bar{z}|^2/\bar{y}) + \frac{|\bar{z}|^2/\bar{y}}{1 + |\bar{z}|^2/\bar{y}}(2 - \frac{|\bar{z}|^2}{|z|^2} - \frac{y}{\bar{y}}) \\ \geq& \ln(1 + |\bar{z}|^2/\bar{y}) \\ &+ \frac{|\bar{z}|^2/\bar{y}}{1 + |\bar{z}|^2/\bar{y}}(2 - \frac{|\bar{z}|^2}{2\Re\{\bar{z}^* z\} - |\bar{z}|^2} - \frac{y}{\bar{y}}),\end{aligned} \tag{A.2}$$

over the trust region

$$2\Re\{\bar{z}^* z\} - |\bar{z}|^2 > 0. \tag{A.3}$$

Setting $y = \bar{y} = \sigma^2$ in (A.2) leads to

$$\begin{aligned}\ln(1 + |z|^2/\sigma^2) \geq& \ln(1 + |\bar{z}|^2/\sigma^2) \\ &+ \frac{|\bar{z}|^2/\sigma^2}{1 + |\bar{z}|^2/\sigma^2}(1 - \frac{|\bar{z}|^2}{2\Re\{\bar{z}^* z\} - |\bar{z}|^2}),\end{aligned} \tag{A.4}$$

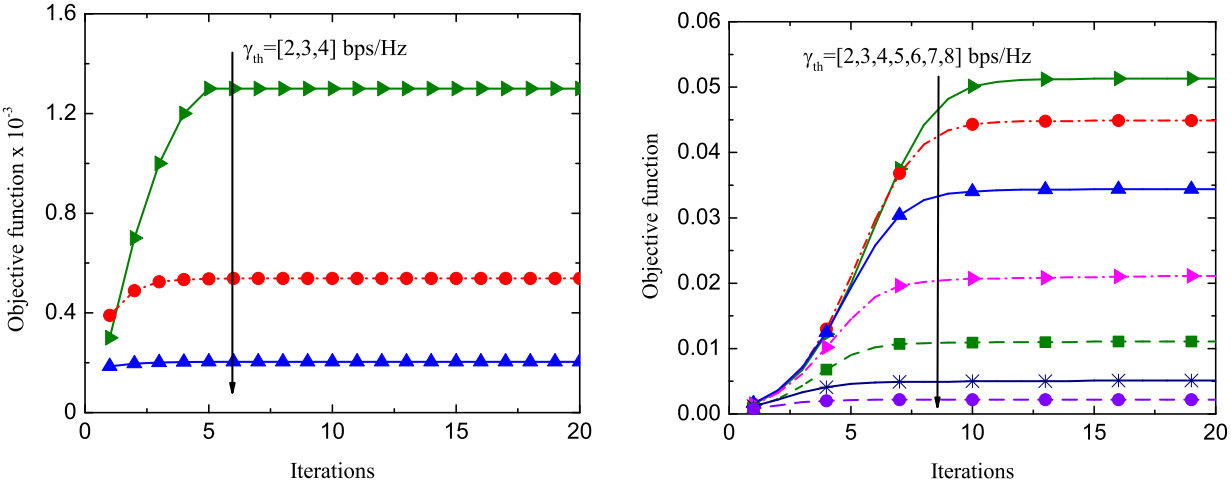
over the trust region (A.3).

Also

$$\begin{aligned}&\ln\left(1 + \frac{|z|^2}{|y|^2 + \sigma^2}\right) \\ &\leq \ln(1 + \frac{|\bar{z}|^2}{|\bar{y}|^2 + \sigma^2}) + \left(1 + \frac{|\bar{z}|^2}{|\bar{y}|^2 + \sigma^2}\right)^{-1} \\ &\quad \times \left(\frac{|z|^2}{|y|^2 + \sigma^2} - \frac{|\bar{z}|^2}{|\bar{y}|^2 + \sigma^2}\right) \\ &\leq \ln(1 + \frac{|\bar{z}|^2}{|\bar{y}|^2 + \sigma^2}) + \left(1 + \frac{|\bar{z}|^2}{|\bar{y}|^2 + \sigma^2}\right)^{-1} \\ &\quad \times \left(\frac{|z|^2}{2\Re\{\bar{y}^* y\} - |\bar{y}|^2 + \sigma^2} - \frac{|\bar{z}|^2}{|\bar{y}|^2 + \sigma^2}\right)\end{aligned} \tag{A.5}$$

over the trust region

$$2\Re\{\bar{y}^* y\} - |\bar{y}|^2 > 0. \tag{A.6}$$



(a) Algorithm 1.

(b) Algorithm 2.

Fig. 8. Convergence behavior of the proposed algorithms.

Particularly,

$$\ln\left(1+\frac{|z|^2}{\sigma^2}\right) \leq \ln(1+\frac{|\bar{z}|^2}{\sigma^2}) + \left(1+\frac{|\bar{z}|^2}{\sigma^2}\right)^{-1}\left(\frac{|z|^2}{\sigma^2}-\frac{|\bar{z}|^2}{\sigma^2}\right). \quad \text{(A.7)}$$

REFERENCES

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019, White Paper*, Feb. 2015.

[2] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[3] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 26–36, Jul. 2012.

[4] H. Zhou, M. Tao, E. Chen, and W. Yu, "Content-centric multicast beamforming in cache-enabled cloud radio access networks," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Dec. 2015, pp. 1–6.

[5] A. H. Phan, H. D. Tuan, H. H. Kha, and D. T. Ngo, "Nonsmooth optimization for efficient beamforming in cognitive radio multicast transmission," *IEEE Trans. Signal Process.*, vol. 60, pp. 2941–2951, Jun. 2012.

[6] S. Schwarz and M. Rupp, "Transmit optimization for the MISO multicast interference channel," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4936–4949, Dec. 2015.

[7] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Wireless backhauling of 5G small cells: challenges and solution approaches," *IEEE Wirel. Commun.*, vol. 22, no. 5, pp. 22–31, Oct. 2015.

[8] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[9] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "In-network caching and content placement in cooperative small cell networks," in *Proc. Int. Conf. 5G Ubiq. Connect. (5GU)*, Nov. 2014, pp. 128–133.

[10] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[11] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[12] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 9, pp. 6284–6297, Sept. 2016.

[13] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 1, pp. 250–264, Jan. 2017.

[14] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, and M. Latva-Aho, "Ultra dense small cell networks: Turning density into energy efficiency," *IEEE J. Selec. Areas Commun.*, vol. 34, no. 5, pp. 1267–1280, May 2016.

[15] J. Liu and S. Sun, "Energy efficiency analysis of dense small cell networks with caching at base stations," in *Prod. IEEE Int. Conf. Computer and Commun. (ICCC)*, Oct. 2016, pp. 2944–2948.

[16] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[17] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 9, pp. 6118–6131, Sept. 2016.

[18] X. Ge, S. Tu, G. Mao, C.-X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[19] S. F. Yunas, M. Valkama, and J. Niemelä, "Spectral and energy efficiency of ultra-dense networks under different deployment strategies," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 90–100, Jan. 2015.

[20] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Info. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[21] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Info. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.

[22] J. S. P. Roig, D. Gündüz, and F. Tosato, "Interference networks with caches at both ends," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[23] J. Kakar, A. Alameer, A. Chaaban, A. Sezgin, and A. Paulraj, "Cache-assisted broadcast-relay wireless networks: A delivery-time cache-memory tradeoff," *https://arXiv:1803.04058*.

[24] O. Somekh, O. Simeone, Y. Bar-Ness, A. M. Haimovich, and S. Shamai, "Cooperative multicell zero-forcing beamforming in cellular downlink channels," *IEEE Trans. Info. Theory*, vol. 55, no. 7, pp. 3206–3219, 2009.

[25] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wirel. Commun.*, vol. 8, no. 4, pp. 1910–1921, 2009.

[26] F. Zhuang and V. K. N. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684–693, 2014.

[27] L. D. Nguyen, H. D. Tuan, and T. Q. Duong, "Energy-efficient signalling in QoS constrained heterogeneous networks," *IEEE Access*, vol. 4, pp. 7958–7966, Nov. 2016.

[28] B. Azari, O. Simeone, U. Spagnolini, and A. M. Tulino, "Hypergraph-based analysis of clustered co-operative beamforming with application to edge caching," *IEEE Wirel. Commun. Letters*, vol. 5, no. 1, pp. 84–87, Jan. 2016.

[29] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, vol. 1, Mar. 1999, pp. 126–134.
[30] H. Tuy, *Convex Analysis and Global Optimization (Second edition)*. Springer, 2017.
[31] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operation Research*, vol. 26, no. 4, pp. 681–683, 1978.
[32] E. Bjornson, M. Kountouris, and M. Debbah, "Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination," in *Proc. IEEE ICT*, May 2013, pp. 1–5.
[33] J. Tang, D. K. C. So, E. Alsusa, K. A. Hamdi, and A. Shojaeifard, "Resource allocation for energy efficiency optimization in heterogeneous networks," *IEEE J. on Selected Areas in Commun.*, vol. 33, no. 10, pp. 2104–2117, Oct 2015.
[34] A. A. Nasir, H. D. Tuan, D. T. Ngo, T. Q. Duong, and H. V. Poor, "Beamforming design for wireless information and power transfer systems: Receive power-splitting vs transmit time-switching," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 876–889, 2017.

Trung Q. Duong (S'05, M'12, SM'13) received his Ph.D. degree in Telecommunications Systems from Blekinge Institute of Technology (BTH), Sweden in 2012. Currently, he is with Queen's University Belfast (UK), where he was a Lecturer (Assistant Professor) from 2013 to 2017 and a Reader (Associate Professor) from 2018. His current research interests include Internet of Things (IoT), wireless communications, molecular communications, and signal processing. He is the author or co-author of 290 technical papers published in scientific journals (165 articles) and presented at international conferences (125 papers).

Dr. Duong currently serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IET COMMUNICATIONS, and a Lead Senior Editor for IEEE COMMUNICATIONS LETTERS. He was awarded the Best Paper Award at the IEEE Vehicular Technology Conference (VTC-Spring) in 2013, IEEE International Conference on Communications (ICC) 2014, IEEE Global Communications Conference (GLOBECOM) 2016, and IEEE Digital Signal Processing Conference (DSP) 2017. He is the recipient of prestigious Royal Academy of Engineering Research Fellowship (2016-2021) and has won a prestigious Newton Prize 2017.

Huy T. Nguyen received the B.S. degree (Hons.) in Computer Science and Engineering from HCMC University of Technology, Vietnam, in 2013. He received the M.S. degree from the Department of Information and Communication System, Inje University, South Korea in 2016, where he is currently working toward a Ph.D. degree. He was a visiting student with Queens University of Belfast, U.K. in 2017 and internship Ph.D student in National Institute of Information and Communications Technology (NICT), Japan in 2018. His research interests include performance analysis and optimization techniques in signal processing for wireless communications.

H. Vincent Poor (S'72, M'77, SM'82, F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering. From 2006 until 2016, he served as Dean of Princetons School of Engineering and Applied Science. He has also held visiting appointments at several other institutions, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the recent book Information Theoretic Security and Privacy of Information Systems (Cambridge University Press, 2017).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences, the Royal Society and other national and international academies. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, Honorary Professorships from Peking University and Tsinghua University, both conferred in 2017, and a D.Sc. honoris causa from Syracuse University, awarded in 2017.

Hoang Duong Tuan received the Diploma (Hons.) and Ph.D. degrees in applied mathematics from Odessa State University, Ukraine, in 1987 and 1991, respectively. He spent nine academic years in Japan as an Assistant Professor in the Department of Electronic-Mechanical Engineering, Nagoya University, from 1994 to 1999, and then as an Associate Professor in the Department of Electrical and Computer Engineering, Toyota Technological Institute, Nagoya, from 1999 to 2003. He was a Professor with the School of Electrical Engineering and Telecommunications, University of New South Wales, from 2003 to 2011. He is currently a Professor with the School of Electrical and Data Engineering and a Core Member of the Global Big Data Technologies Centre, University of Technology Sydney. He has been involved in research with the areas of optimization, control, signal processing, wireless communication, and biomedical engineering for more than 20 years.

Won-Joo Hwang (M'03, SM'17) received the Ph.D. Degree in Information Systems Engineering from Osaka University Japan in 2002. He received his B.S. and M.S. degree in Computer Engineering from Pusan National University, Pusan, Republic of Korea, in 1998 and 2000. He is now a full professor at Inje University, Gyeongnam, Republic of Korea. His research interests are in network optimization and cross layer design. Dr. Hwang is a member of the IEICE and IEEE.